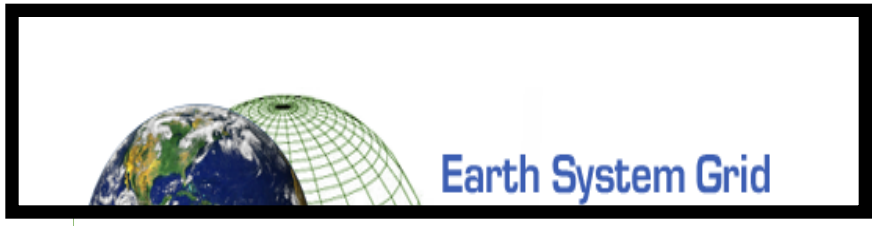
Enabling Climate Model Intercomparison Projects with CDAT and the Earth System Grid



Bob Drach and Dean Williams

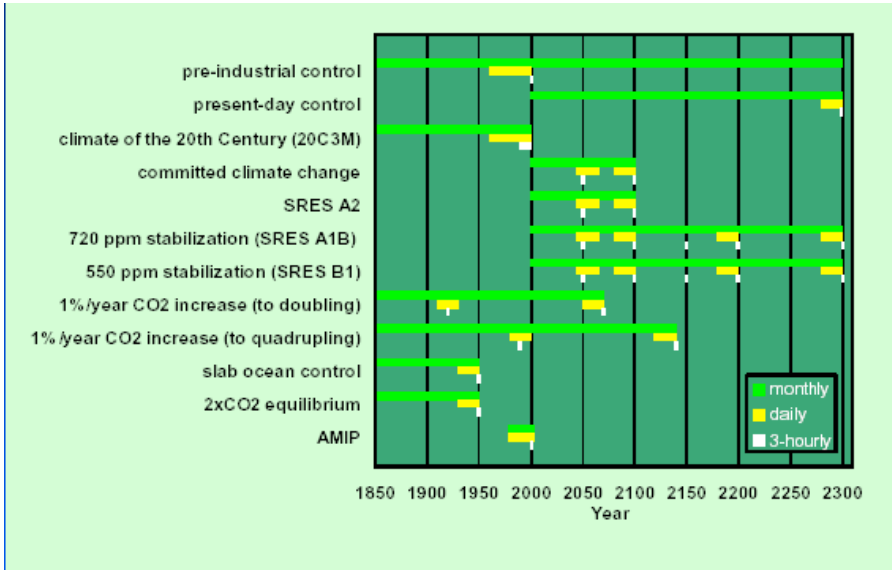
PCMDI/LLNL

WCRP/CMIP3 Multi-Model Database

- Supports the IPCC Fourth Assessment Report (4AR)
- IPCC: Intergovernmental Panel on Climate Change
  - Issues a report every six years
  - "A comprehensive and rigorous picture of the global present state of knowledge of climate change."
  - Huge undertaking: 3000+ reviewers and authors
  - 4AR Volume I will be released February 2, 2007
- Working Group 1 focuses on the physical climate system: atmosphere, land surface, ocean, and sea ice
  - World Climate Research Programme coordinated the experimental design
  - 12 scenarios, 20+ coupled atmosphere/ocean climate models
- Database hosted at Program for Climate Model Diagnosis and Intercomparison / LLNL
- Software developed by Earth System Grid II project

Scenarios, Models

- Scenario is the specification of one experiment
  - Some groups submitted ensembles of runs
- Temporal frequency of datasets: yearly, monthly, daily, 3-hourly



Modeling Center	Model
BCCR, Norway	BCM 2.0
CCCma, Canada	CGCM3.1
CCSR/NIES/FRCGC (hi-res), Japan	MIROC3.2 (hi-res)
CCSR/NIES/FRCGC (med-res), Japan	MIROC3.2 (med-res)
CNRM, France	CNRM-CM3
CSIRO, Australia	CSIRO Mk3.0
GFDL (CM2.0), USA	GFDL_CM2.0
GFDL (CM2.1), USA	GFDL_CM2.1
GISS (C4x3), USA	C4x3
GISS (Model E-H), USA	Model E-HYCOM
GISS (Model E-R), USA	Model E-Russell

Modeling Center	Model
IAP, China	FGOALS1.0_g
INM, Russia	INMCM3.0
IPSL, France	IPSL-CM4
MPI, Germany	ECHAM 5 / MPI-OM
MRI, Japan	MRI-CGCM2.3.2a
NCAR (CCSM3), USA	CCSM3.0
NCAR (PCM1), USA	PCM1
NCC, China	CSM T63 (Temporal)
UKMO (HadCM3), UK	HadCM3
UKMO (HadGEM1), UK	HadGEM1

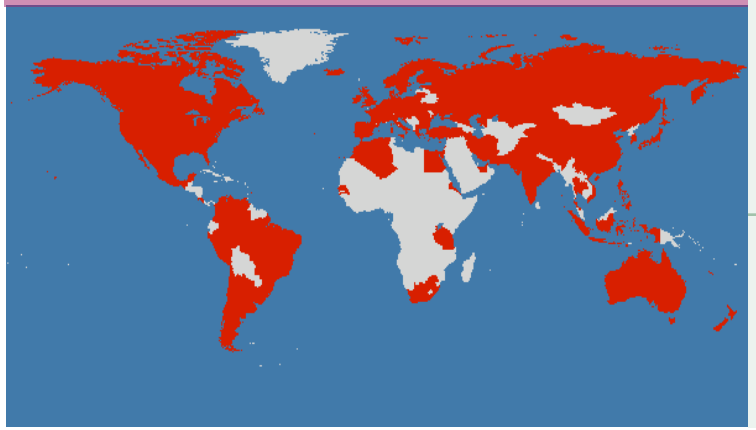
ESG facts and figures

Earth System Grid

ESG Objective

To support the infrastructural needs of the national and international climate community, ESG is providing crucial technology to securely access, monitor, catalog, transport, and distribute data in today's Grid computing environment.

Worldwide ESG user base



WCRP/CMIP3 Portal

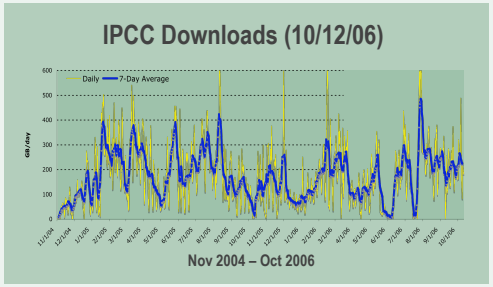
31 TB of data at the PCMDI site location

- 71,900 files
- Generated by a modeling campaign coordinated by the Intergovernmental Panel on Climate Change
- Model data from 12 countries

950 registered users

Downloads to date

- 150 TB
- 639,000 files
- 300 GB/day (average)



200+ scientific papers published to date based on analysis of WCRP/CMIP3 IPCC AR4 data

AR5 database will use ESG-CET

Earth System Grid

- Building on the success of the WCRP/CMIP3 ESG data portal.
- For IPCC AR5, data volumes will be 10x larger
  - How best to collect and distribute data on a much larger scale?
- Models are becoming more complex (e.g., carbon cycle, dynamic vegetation, etc.)
- How to make information understandable to end-users so that they can interpret the data correctly?
  - Users may be part of WG2 (impacts), WG3 (mitigation)
- More ambitious community modeling projects (>300 TBs) demand:
  - Distributed data environment - avoid unnecessary data movement.
  - Client and Server-side analysis and visualization tools (subsetting, concatenating, regridding, filtering, …)
- Testbed needed by late 2008 – early 2009

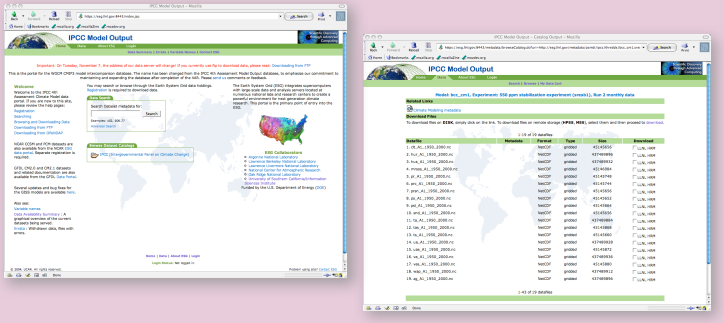
Providing climate scientists with access to simulation results needed for their research

Earth System Grid

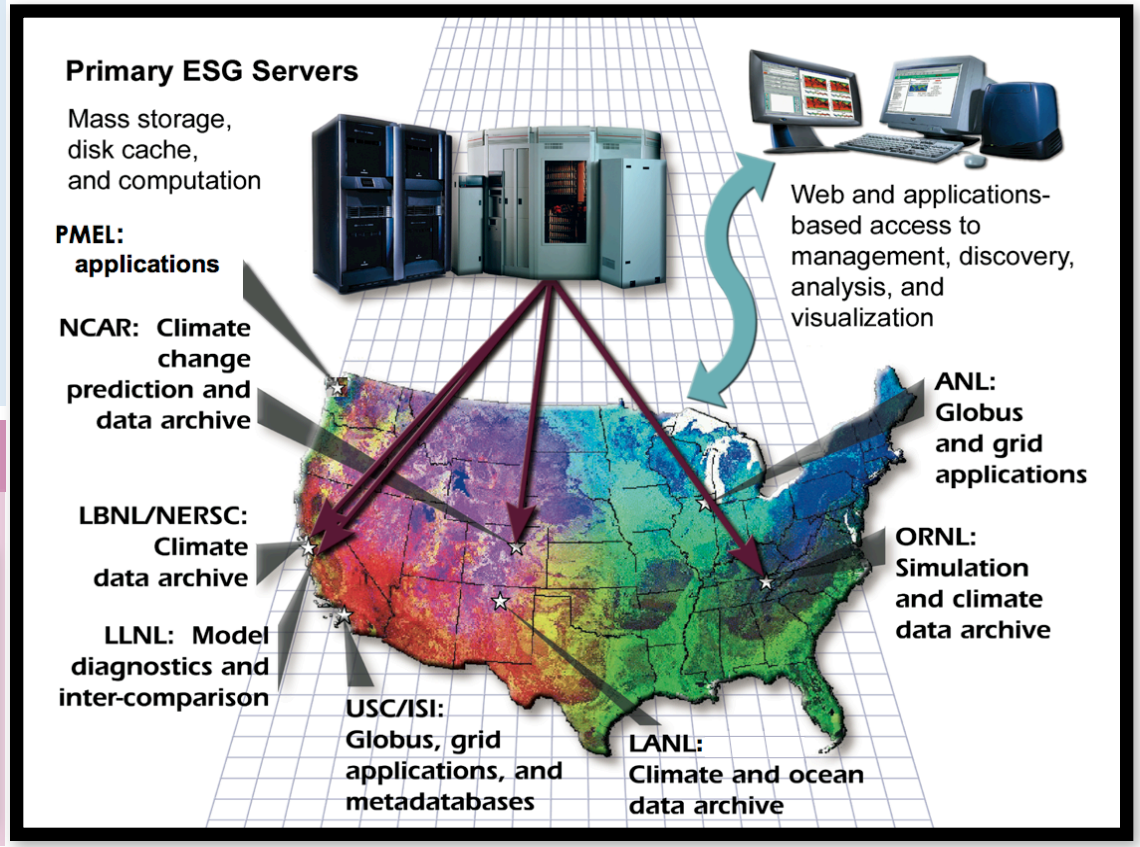
ESG Goal

- Very large distributed data archives
  - ➢Easy federation of sites
  - ➢Across the US and around the world
- "Virtual Datasets" created through subsetting and aggregation
- Metadata-based search and discovery
- Access through browser, analysis tools
- Increased flexibility and robustness
- Server-side analysis

http://www-pcmdi.llnl.gov



Current ESG Sites



Evolving ESG for the future

Earth System Grid

ESG Data System Evolution

2006

Central database

- Centralized curated data archive
- Time aggregation (virtual datasets)
- Distribution by file transport
- No ESG responsibility for analysis
- Shopping-cart-oriented web portal
- ESG connection to desktop analysis tools (i.e., CDAT and CDAT-LAS)

Early 2009

Testbed data sharing

- Federated metadata
- Federated portals
- Quick look server-side analysis with CDAT
- Location independence
- Distributed aggregation
- Manual publishing

2011

Full data sharing

- Full suite of server-side analysis with CDAT
- Model/observation comparison
- ESG embedded into desktop productivity tools with CDAT
- GIS integration
- Notification services
- Models integrated with ESG

CCSM AR4

ESG Data Archive

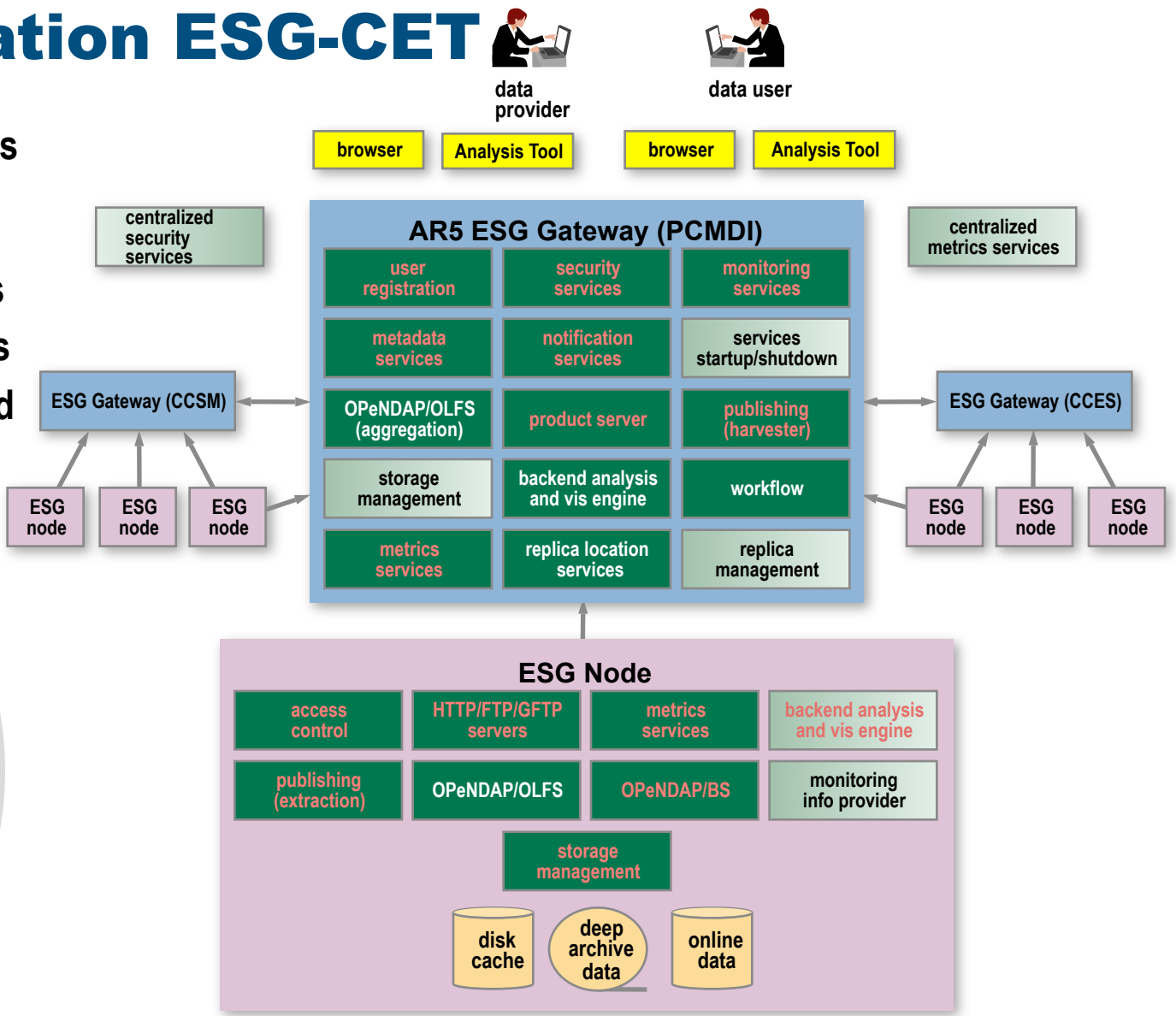
CCSM, AR5, biogeochemistry, in-situ data, …

Terabytes

Petabytes

Architecture of the next-generation ESG-CET

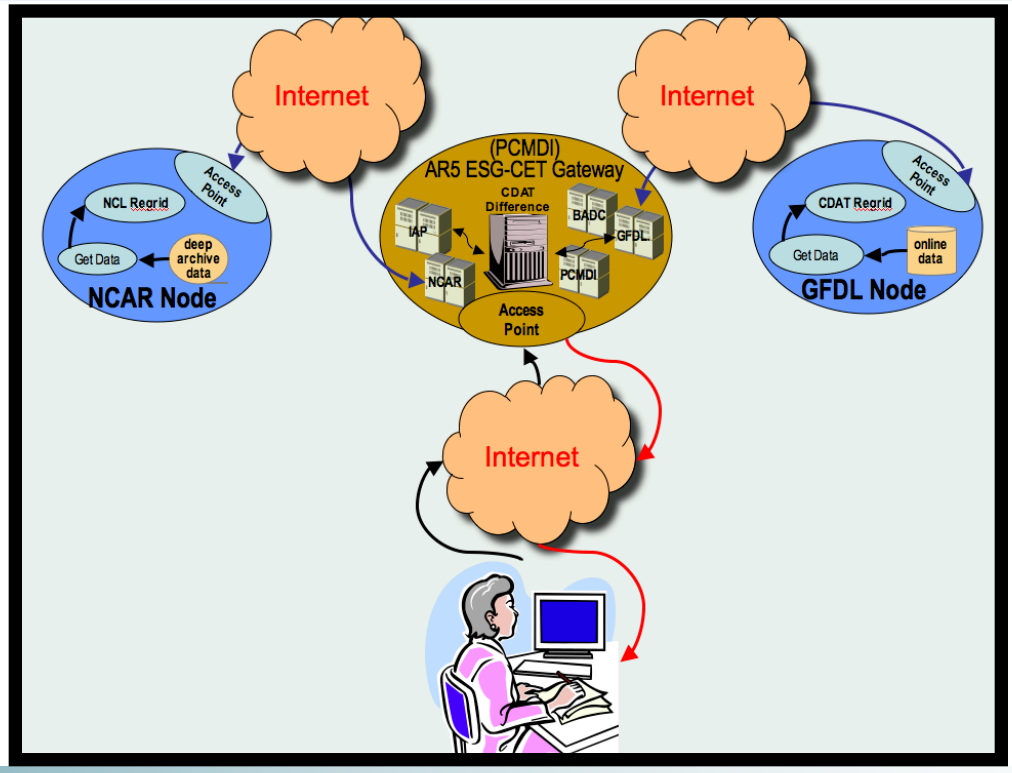
- Very large data archives (>300TB)
- Broader geographical distribution of archives
- Easy federation of sites
- Increased flexibility and robustness



Simple intercomparison use case scenario

Current Scenario	Future Scenario

• Browse PCMDI's centralized database
• Download data
• Organize data on local site
• Regrid data at local site
• Perform diagnostics
• Produces results
	• Search, browse and discover distributed data
	• Remote site
	  ➢ Request data
	  ➢ Regrids
	  ➢ Diagnostics
	• ESG returns results



Climate Data Analysis Tools: data analysis, and visualization for intercomparison research

CDAT Goal

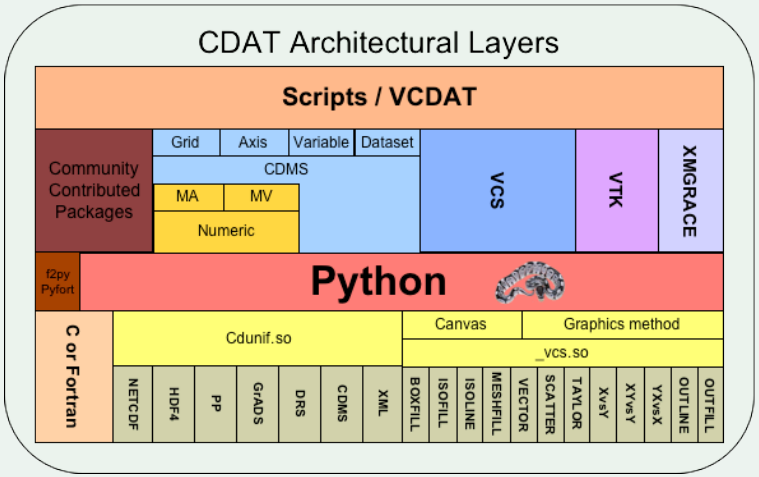
Address the challenges of enabling data management, discovery, access, and advanced data analysis for climate model diagnosis and intercomparison research.

Typical usage examples of CDAT

- Calculate a long-term average
- Define wind-speed from u- and v-components
- Subset a dataset, selecting a spatiotemporal region
- Aggregate 1000s of files into a small XML file
- Generate a Hovmoller plot

What is CDAT?

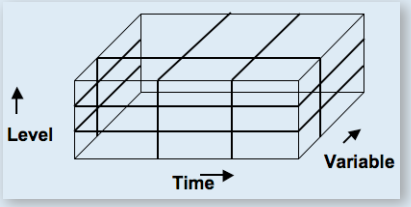
- CDAT *IS* Python!
- Designed for climate science data
- Scriptable
- Freely available



Examples of CDAT analysis, aggregation

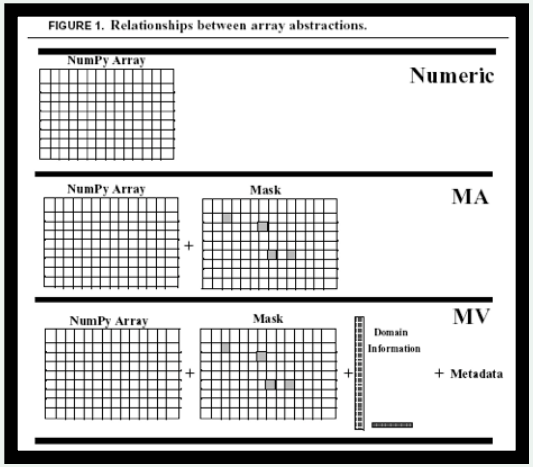
CDSCAN

- Data aggregation: collections of files/datasets are treated as single entities.
- Aspects of aggregation:
  - combining/merging variables,
  - joining variables,
  - new coordinate axes,
  - overlaying/adding metadata,
  - nesting datasets

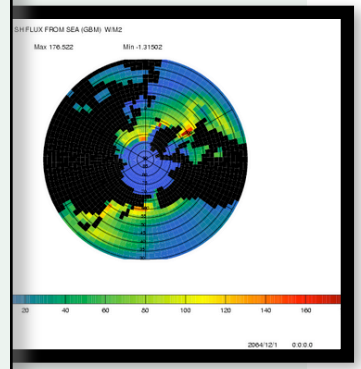


- PCMDI CDAT supports aggregations via the cdscan utility that uses XML representation
- cdscan will analyse the archive for:
  - variable information
  - axis information
  - global (universal) metadata
- Why use cdscan
  - Large datasets described as a grouped entity.
  - No need to know underlying data format.
  - No need to know file-names.
  - Datasets can be sliced in any way the user chooses using logical spatio-temporal selectors rather than loops of programming code.
  - You can use it to improve the metadata of your data files…
- cdscan in action
  - $ cdscan –x monthly_means.xml ./*.nc

MV



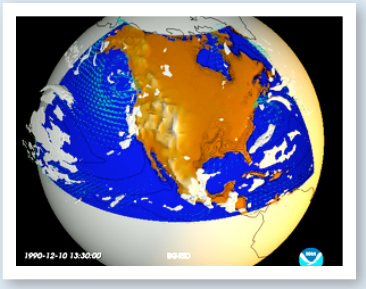

```
>>> import cdms, MV
>>> f_surface = cdms.open('sftlf_ta.nc')
>>> surf = f_surface('sftlf')

# Designate land where "surf" has values
# not equal to 100
>>> land_only = MV.masked_not_equal(surf, 100.)
>>> land_mask = MV.getmask(land_only)

# Now extract a variable from another file
>>> f = cdms.open('ta_1994-1998.nc')
>>> ta = f('ta')

# Apply this mask to retain only land values.
>>> ta_land = cdms.createVariable(ta,
        mask=land_mask, copy=0, id='ta_land')
```

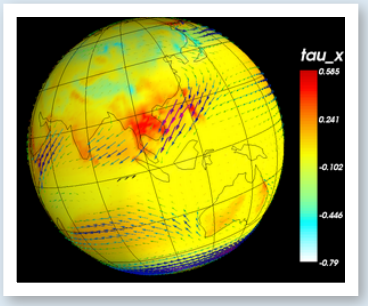


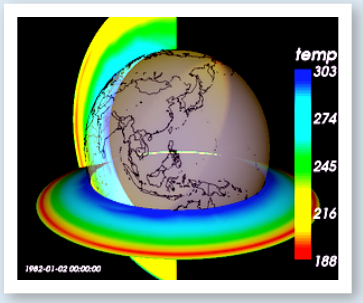
CDAT examples

Ncvtk

Collaboration:
CDAT developers are currently working with Ncvtk developers to make Ncvtk 3D graphics accessible to the CDAT community. Ncvtk is a collection of commonly used 3D visualization methods applied to data on structured lat/lon grids.

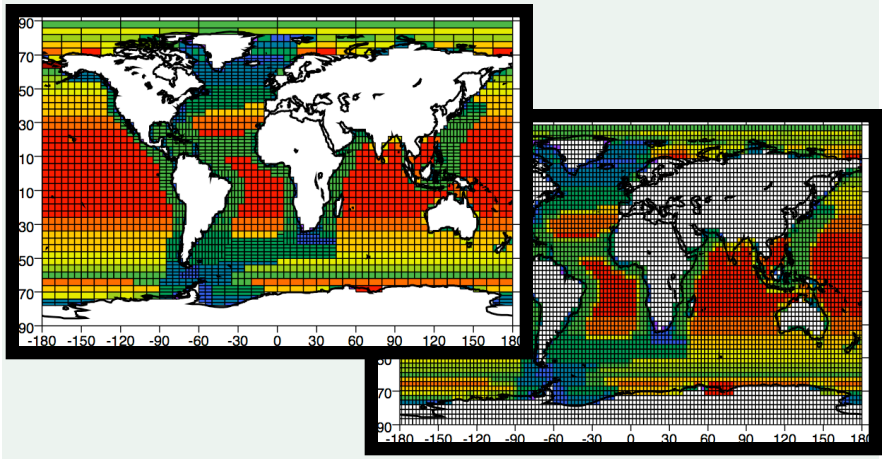




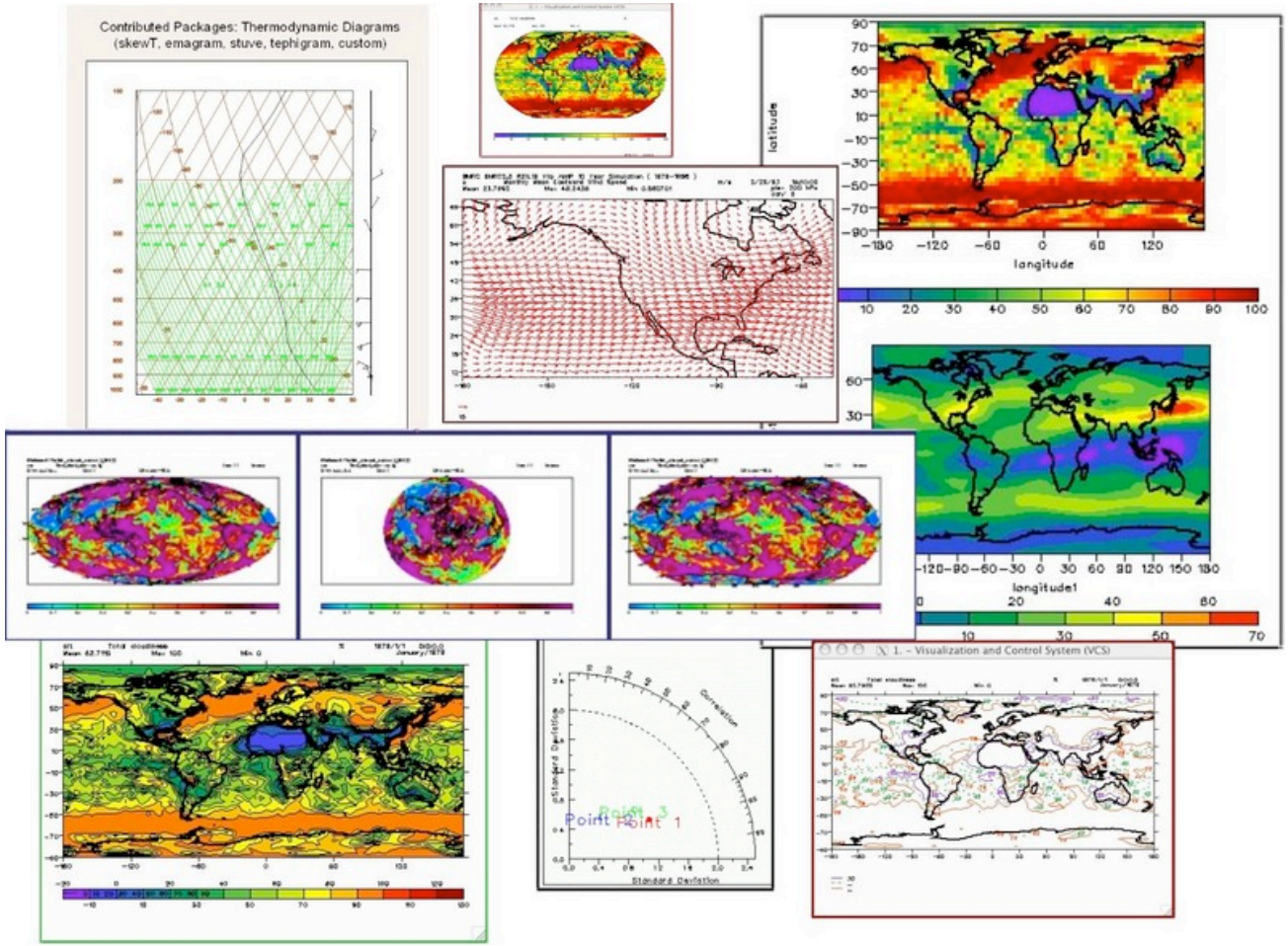


Regridder

```
#!/usr/local/cdat/bin/python
import cdms
from regrid import Regridder
f = cdms.open('temp.nc')
t= f.variables['t']
ingrid = t.getGrid()
outgrid = cdms.createUniformGrid( -90.0, 46, 4.0, 0.0, 72, 5.0)
regridFunc = Regridder(ingrid, outgrid)
newt = regridFunc(t)
import vcs
vcs.init().plot(t)
vcs.init().plot(newt)
```



CDAT examples

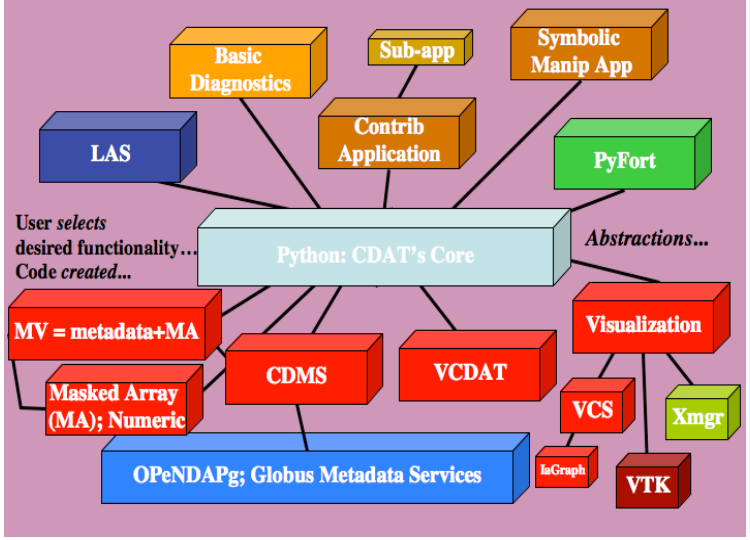


CDAT facts and figures

CDAT Users

- Over 120 mailing list registers
  - Probably 7 to 10 times more casual users
- Mailing list archive: over 1,000 message (~30 per month)
- 912 Downloads since May 19, 2006 for version 4.1
- Improved Documentation

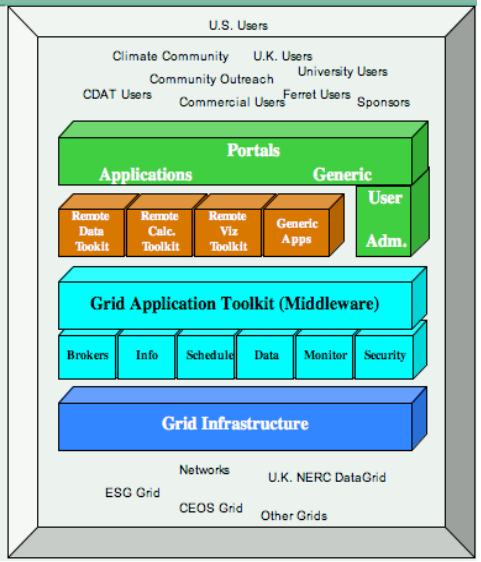
CDAT Core Modules



CDAT Collaborations

Some CDAT development centers:

- British Atmospheric Data Center
- LBNL
- GFDL
- Laboratory of Science of Climate and the Environment (LSCE), FR
- PCMDI
- University of Chicago
- University of Hawaii
- University of Reading, UK



Evolving CDAT into an integrated client technology workplace

CDAT Integrated Analysis Evolution

2006

Community software

- Python based
- Start to finish environment
- Diverse analysis tools
- Languages: C/C++, FORTRAN, Python
- Platforms: Unix, Mac
- VCDAT: discover, learn, and browse with a few clicks
- Connection to ESG-II

Early 2009

Testbed distributed analysis

- Access to shared resources (Web/Grid services)
- Quick look server-side analysis tool for ESG-CET
- Diagnostics specific to AR5
- GFDL Ncvtk 3D visualization
- Web-CDAT: discover, learn, and browse via web browser
- Serving Google Maps and Google Earth Data with CDAT

2011

Full analysis sharing

- Full suite server-side analysis tool for ESG-CET
- ESG-CET embedded into desktop productivity tools (i.e., CDAT)
- GIS integration with CDAT
- SciDAC VACET analysis and visualization collaboration
- Global Organization for Earth System Science Portal (GO-ESSP)

CDMS
Numeric / MV
Genutil / Cdutil
VCS

CDAT Core Modules

CDMS, Numeric, Genutil, Cdutil, Ncvtk, VACET, Diagnostics, ESG

Standalone **Distributed**

Suggestions

- Focus on exploration / elucidation of data (as opposed to purely presentation graphics)
  - Visualization in time
    - Highlight features of interest
  - Comparison of model output / observations
    - Example: Taylor diagrams illustrate improvement in model skill
  - Comparison between models
    - Example: Spaghetti diagrams, portrait diagrams
- Provide a Python interface.
  - Leverage existing data constructs: MaskedArray, MaskedVariable
  - Minimize the number of calls needed to produce analysis / graphics products.